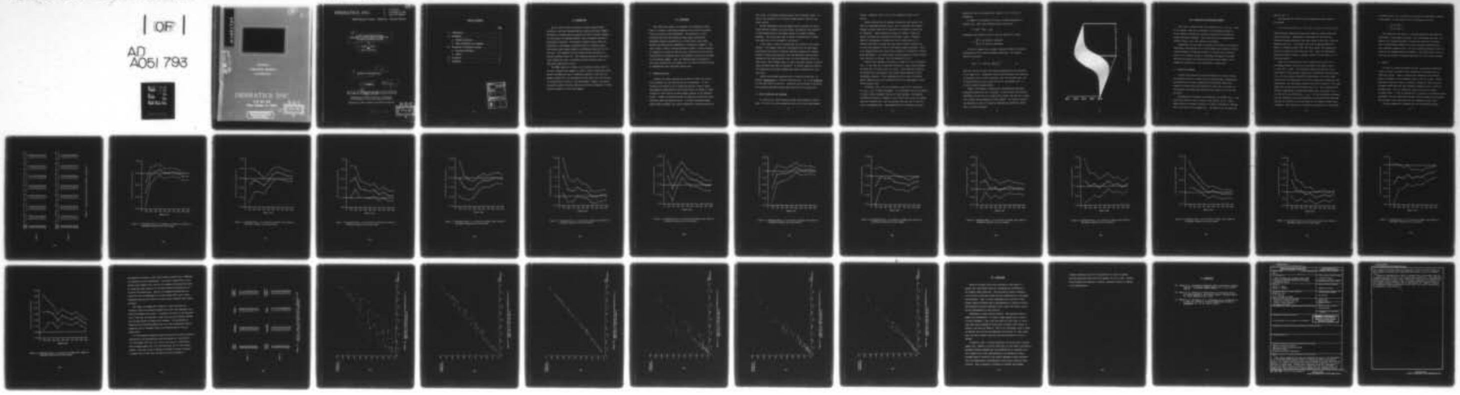
AD-A051 793   DESMATICS INC STATE COLLEGE PA   F/G 6/5
A STUDY OF ESTIMATION ACCURACY WHEN USING A LOGISTIC MODEL FOR --ETC(U)
MAR 78   D E SMITH, R L GARDNER   N00014-74-C-0154
UNCLASSIFIED   TR-102-5   NL

| OF |
AD
A051 793



END
DATE
FILMED
4-78
DDC

12

— STATISTICS —

— OPERATIONS RESEARCH —

— MATHEMATICS —

DESMATICS, INC.

P.O. Box 618
State College, Pa. 16801

(12)

DESMATICS, INC.

P. O. Box 618
State College, Pa. 16801
Phone: (814) 238-9621

Applied Research in Statistics - Mathematics - Operations Research

(6)

A STUDY OF ESTIMATION ACCURACY WHEN USING A LOGISTIC MODEL FOR PREDICTION OF IMPACT ACCELERATION INJURY

by

(10) Dennis E. Smith
[redacted]
Robert L. Gardner

(9) TECHNICAL REPORT NO. (14) TR- 102-5

(11) Mar[redacted]978 (12) 49p.

(15)

This study was supported by the Office of Naval Research
under Contract No. N00014-74-C-0154, Task No. NR 207-037

Reproduction in whole or in part is permitted
for any purpose of the United States Government

Approved for public release; distribution unlimited

391 156

TABLE OF CONTENTS

I. INTRODUCTION

The U.S. Navy's impact acceleration research program being conducted by the Naval Aerospace Medical Research Laboratory (NAMRL) Detachment is accumulating an extensive data base on dynamic response, for both humans and subhuman primates, under stringently controlled experimental conditions. This wealth of empirical data offers the possibility of developing a statistical model of head/neck impact acceleration injury based primarily on information obtained from the data base. The framework of such a model has been discussed in a previous technical report [2]. The following sections of the current report address the topic of estimation accuracy when this type of empirically-based model is used.

The NAMRL research is focused on the head/neck system, which is the most vulnerable body segment in terms of impact acceleration injury. Because the NAMRL data base is comprised primarily of head and neck dynamic response data, the discussion in this report will be restricted to consideration of injury in that body segment. However, the general procedures proposed for model construction should be adaptable to injury prediction models for other body segments.

II. BACKGROUND

Most head injury models, of necessity, are mathematical models based on a number of underlying assumptions about head/neck movement, forces, and overall injury mechanisms. The existence of a well-controlled data base from animal and human impact acceleration experiments permits consideration of a probabilistic injury model derived from empirical data embedded in a statistical framework. This type of model, unlike a standard mathematical model, is based primarily on information contained in observed data, rather than on that derived from theoretical assumptions about the mechanical structure and dynamics of the head/neck segment. Thus, the modeling approach discussed in this report should offer new insights into the injury prediction problem by complementing those approaches usually used.

A. PROBLEM DISCUSSION

Consider the impact acceleration situation in which the torso is well-restrained, but the head and neck are unrestrained. In this situation the problem is one of predicting whether a human of given anthropometric characteristics will sustain injury if exposed to impact acceleration which results in given dynamic response of the head/neck system. A number of difficulties must be overcome to develop an injury prediction model from empirical data. If enough instrumented human subjects were available, and could be subjected to various acceleration time traces, a reasonable prediction model would eventually result. Of course, this procedure is not possible--human subjects cannot be purposely injured.

Because experiments involving humans cannot be planned for potentially injurious regions of the data space, any empirical data gathered in those regions must be from human analogs (for example, subhuman primates). The results must then be scaled or extrapolated to humans. This topic will not be discussed in this report.

In any event, it must be realized that the situation is not deterministic. For example, even with a restrained torso, the same impact acceleration will result in different head/neck response (for example, because of initial head position), and even apparently identical head response for the same person may result in injury sometimes and not at other times. This binomial aspect of injury occurrence defines a discrete random variable which must be considered. To further complicate matters, the acceleration and dynamic response data under consideration is time trace data.

Despite the problems mentioned here, it should be noted that, in general, construction of a model is relatively easy. It is the validation of that model which is difficult. Validation may be defined as establishing acceptable agreement between model predictions and observed data.

B. MODEL ASSUMPTIONS AND FRAMEWORK

To construct any injury prediction model, some assumptions must be made. The trick is to make assumptions which are at least approximately correct. Hopefully, this is true of the assumptions made in this section.

Because dealing with the complete acceleration time traces of the head is an impossible analytic task, a set of univariate head dynamic response variables which may be expected to be related to injury will be considered. Likely candidates include, for example, linear and angular velocities and accelerations (average or peak). Although a number of anthropometric variables can also be postulated, it is probably reasonable to assume that their effect within a species will be minor when compared to that of the dynamic response variables. Thus, it is suggested that only these latter variables be considered in initial model development. At a later stage, the anthropometric variables should prove important in scaling. (See the discussion in [2].)

Exactly which variables to include in a model will not be discussed here. If the total set of potential variables is large, some preliminary screening will, of course, have to be done. Should important variables be excluded from consideration, this should become apparent as model development proceeds. If all experimental data is saved, the primary penalty imposed by such an occurrence would be the requirement for additional analysis time.

In general, then, a set of k variables, which will be denoted by $\underline{x} = (x_1,\ldots,x_k)$, is being considered. It is postulated that the probability of injury is some (unknown) function of these variables. Furthermore, although the function is unknown, it will be near zero in part of $\underline{x}$-space, near one in another part, and will increase from near zero to near one over an intermediate part. Experimentally what is observed in a given situation is only an estimated value (either 0 or 1) of the true probability.

In summary, the probability of injury is being considered as a function of $\underline{x}$. Thus, this probability may be denoted by:

$$P = P(\underline{x}) = P(x_1,\ldots,x_k).$$

Furthermore, the observed value of P will be denoted by y, where:

$$y = \begin{cases} 1 \text{ if an injury is sustained} \\ 0 \text{ if no injury is sustained.} \end{cases}$$

It will be assumed that a logistic function provides a reasonable approximation to the function defining probability. The logistic function is given by:

$$P(\underline{x}) = \{1 + \exp[-(\beta_0 + \sum_1^k \beta_i x_i)]\}^{-1} \qquad (1)$$

When this function is used, all predicted probabilities are restricted to the range (0,1). Furthermore, this function satisfies the conditions of being near zero in a part of $\underline{x}$-space, near one in another part, and increasing from near zero to near one over an intermediate part. It is also tractable computationally.

Figure 1 illustrates a representative approximating probability prediction function for two x variables. As can be seen, the predicted probability is near zero for x_1 small and x_2 small, near one for x_1 large and x_2 large, and intermediate in other regions. The logistic function can represent a variety of shapes by adjusting the coefficient values. Thus, it is quite flexible.

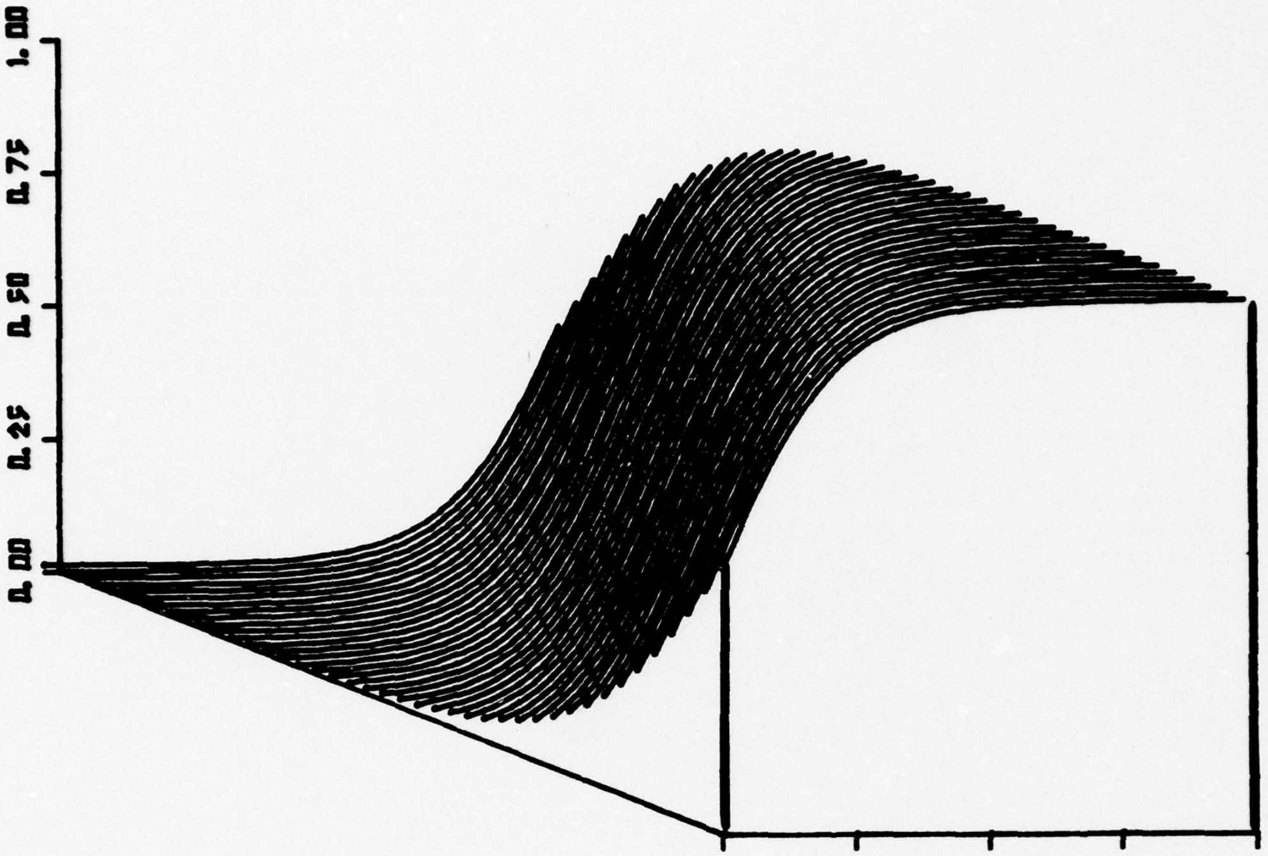


Figure 1: A Representative Approximating Probability Prediction Function for Two Variables

III. EVALUATION OF ESTIMATION ACCURACY

From a set of observed data, the coefficients (i.e., β_0, $\beta_1,\ldots,\beta_k$) of the logistic model may be estimated. The estimation process is fairly complex, involving an iterative procedure which provides the maximum likelihood estimates. This does not pose an insurmountable problem, however, since the computer is available.

Nonetheless, the data input to a model of this kind is of necessity dichotomous, requiring the use of larger samples than required to obtain a desired degree of predictive accuracy if the data were continuous. Thus, it is of central concern to investigate the degree of accuracy which may be expected for predictions derived from the model, and to examine the sensitivity of such predictions to sample size.

A. EVALUATION PROCEDURE

A Monte Carlo simulation study was undertaken to provide information relating accuracy to sample size for selected model configurations. Two specific sets of model parameters were considered, Monte Carlo samples of various sizes were generated for each, and the accuracy of the resultant predictions were evaluated with respect to the true probabilities.

Two models were considered, each employing six variables ($x_1,\ldots,x_6$), which were allowed to take on values in the interval (-1,1). These models, which are hereafter referred to as Model A and Model B, differed only in the value of the parameter β_0. For Model A, $\beta_0 = 0$, while for

Model B, $\beta_0 = -2$.

The remaining six coefficients were assigned the same values in both models:

$$\beta_1 = -0.25,\ \beta_2 = 0.50,\ \beta_3 = -0.75,\ \beta_4 = 1.00,\ \beta_5 = -1.25,\ \beta_6 = 1.50 \ .$$

These particular coefficient values were chosen to produce models with certain properties. Specifically, these models are such that the minimum attainable probability over the $\underline{x}$ region is near zero, while the maximum is near one. In addition, while the average probability for Model A is moderately high (.500) over the $\underline{x}$ region, the average probability for Model B is relatively low (.225). Thus, the observations generated by Model A would, in general, consist of more values of $y = 1$ than would Model B.

Monte Carlo procedures were used to generate two series of ten overlapping samples (one series for each model) with sample sizes of $n = 100, 200, \ldots, 1000$. Each individual sample contained all of the observations in the preceding samples plus an additional 100 observations (i.e., the first sample contained 100 observations, the second sample contained the 100 observations of the first sample plus an additional 100 observations, the third sample contained the 200 observations of the second sample plus an additional 100 observations, etc.). Each observation was defined by generating a uniform random number over the interval (-1,1) for each of the six variables $x_1, \ldots, x_6$. The true probability associated with any observation $\underline{x}$ was then determined by calculating $P(\underline{x})$ from model equation (1) using the true coefficients for the respective model being considered. Each observation was defined as resulting in an "occurrence" or "nonoccurrence" (e.g., an injury or noninjury) by generating a uniform random number r in the interval (0,1) and defining y such that

$$y = \begin{cases} 1 \text{ if } P(\underline{x}) > r \\ 0 \text{ if } P(\underline{x}) \leq r \, . \end{cases}$$

Each sample was then input to a computer program for estimating the coefficients of the logistic function. Two such programs were used. One, adapted from a program developed at the National Institutes of Health, uses the Walker-Duncan method [3] to obtain estimated coefficients and their estimated standard errors. The other program developed by Jones [1], solves for the maximum likelihood estimates. Both programs were found to yield the same results; the primary difference is in the output provided.

B. RESULTS

In order to evaluate estimation accuracy, the estimated coefficient values for Model A and Model B were compared with the respective true coefficient values. Figure 2 provides this comparison, and indicates general convergence of the estimated values to the true values. Convergence may be seen more clearly in Figures 3 through 16, which graph the estimated coefficients and their estimated standard error as a function of sample size. The true coefficient values are shown as straight horizontal lines. It may be observed that, in general, the estimated values approximate the true values more and more closely as sample size increases. This provides a clear indication that the estimation process works.

Although comparison of estimated and true coefficients provides

	Sample Size	$\beta_0 = 0.00$	$\beta_1 = -0.25$	$\beta_2 = 0.50$	$\beta_3 = -0.75$	$\beta_4 = 1.00$	$\beta_5 = -1.25$	$\beta_6 = 1.50$
Model A	100	-0.43	-0.43	0.92	-0.78	1.53	-1.03	0.89
	200	-0.06	-0.31	1.02	-1.00	1.16	-1.36	1.36
	300	0.03	-0.35	0.73	-1.07	1.11	-1.05	1.39
	400	0.07	-0.45	0.70	-1.06	1.22	-0.86	1.40
	500	-0.01	-0.30	0.66	-0.91	1.14	-1.02	1.50
	600	0.01	-0.14	0.70	-0.82	1.12	-1.12	1.63
	700	0.02	-0.09	0.58	-0.85	0.97	-1.15	1.52
	800	-0.02	-0.14	0.64	-0.79	0.96	-1.21	1.53
	900	-0.08	-0.19	0.57	-0.76	1.00	-1.28	1.54
	1000	-0.07	-0.20	0.56	-0.80	1.06	-1.27	1.49

	Sample Size	$\beta_0 = -2.00$	$\beta_1 = -0.25$	$\beta_2 = 0.50$	$\beta_3 = -0.75$	$\beta_4 = 1.00$	$\beta_5 = -1.25$	$\beta_6 = 1.50$
Model B	100	-2.27	-0.06	0.64	-0.00	1.43	-1.85	2.31
	200	-2.05	0.01	0.50	-0.22	1.05	-1.63	2.34
	300	-2.08	-0.24	0.55	-0.40	1.05	-1.67	2.26
	400	-2.07	-0.20	0.44	-0.61	0.91	-1.56	1.95
	500	-2.16	-0.30	0.54	-0.76	0.95	-1.66	2.17
	600	-2.19	-0.23	0.63	-0.70	0.80	-1.62	2.07
	700	-2.21	-0.18	0.52	-0.77	0.83	-1.64	2.00
	800	-2.25	-0.23	0.55	-0.86	0.90	-1.55	2.09
	900	-2.17	-0.23	0.46	-0.82	0.83	-1.49	1.95
	1000	-2.15	-0.26	0.54	-0.85	0.91	-1.45	1.81

Figure 2: Estimated Coefficients for Model A and Model B.

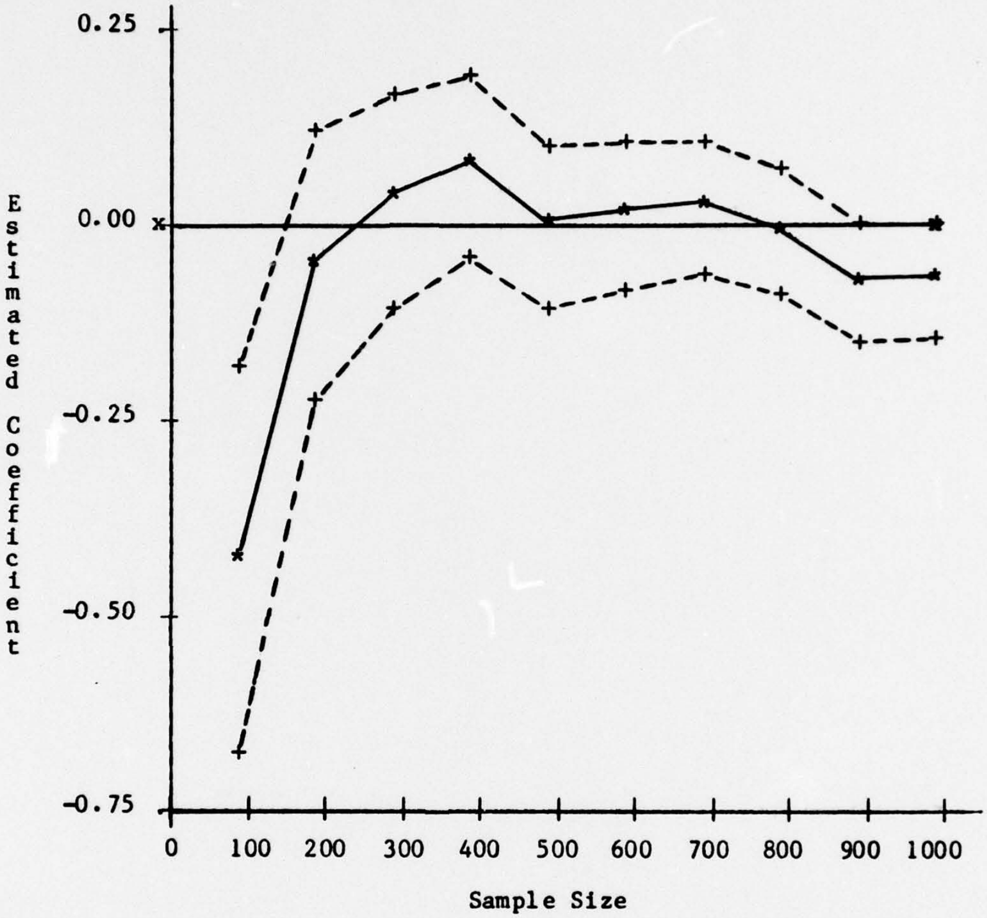


Figure 3: Estimated Beta-0 as a Function of Sample Size (Model A)
(Estimated Standard Error Also Shown)

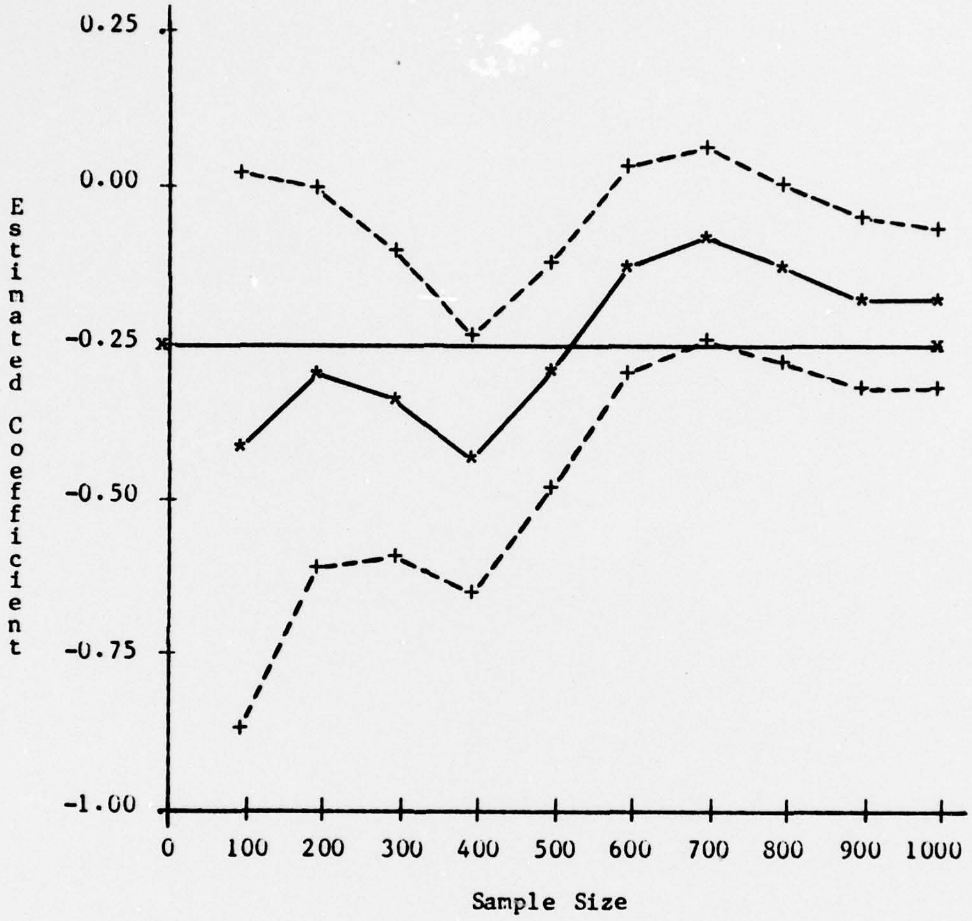


Figure 4: Estimated Beta-1 as a Function of Sample Size (Model A)
(Estimated Standard Error Also Shown)

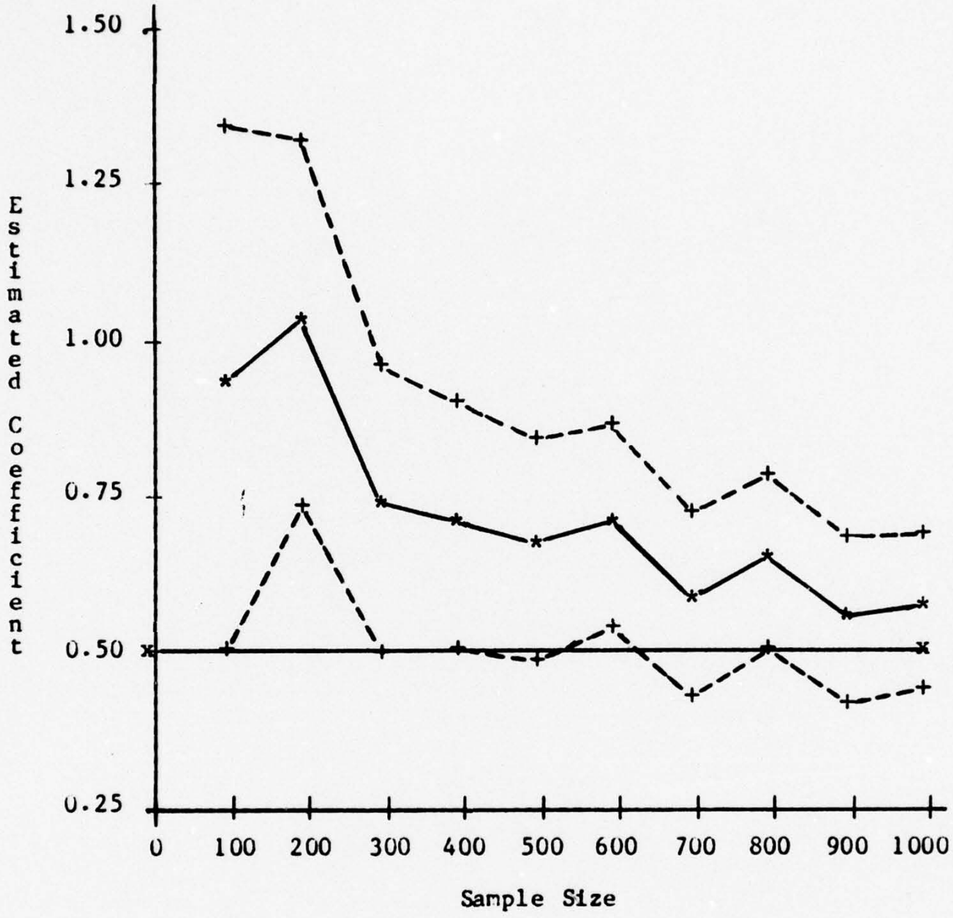


Figure 5: Estimated Beta-2 as a Function of Sample Size (Model A)
(Estimated Standard Error Also Shown)

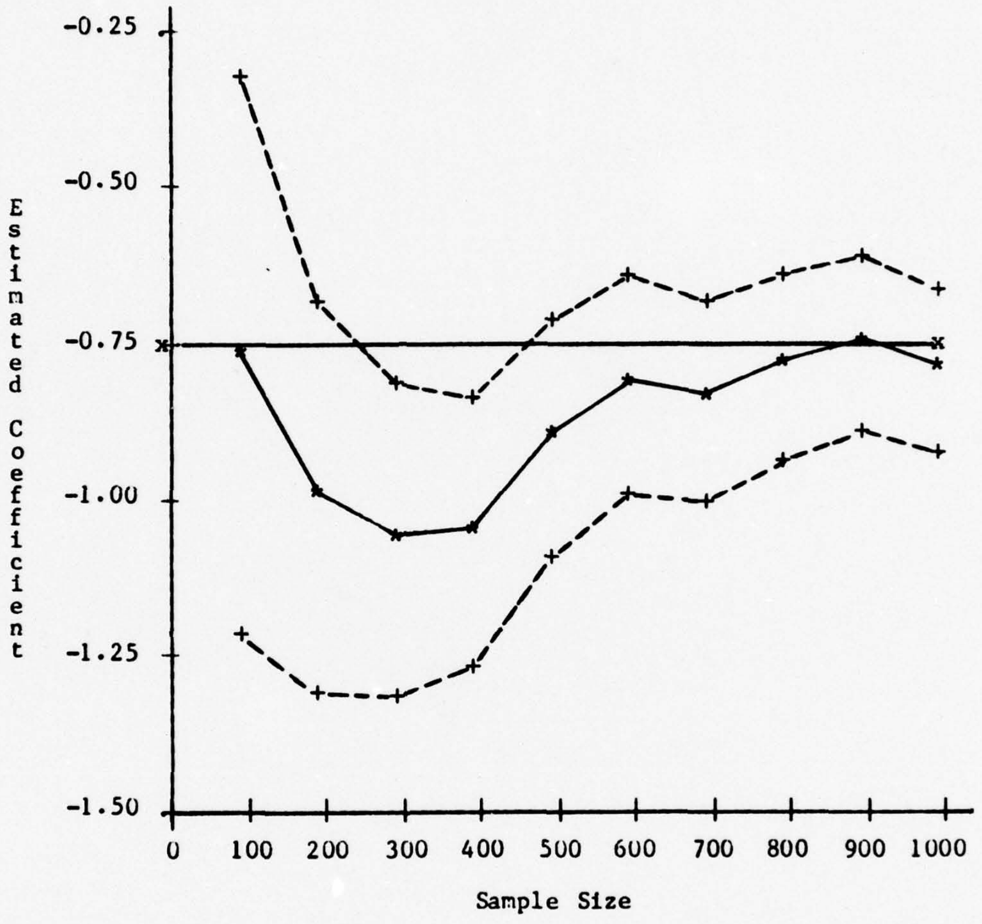


Figure 6: Estimated Beta-3 as a Function of Sample Size (Model A)
(Estimated Standard Error Also Shown)

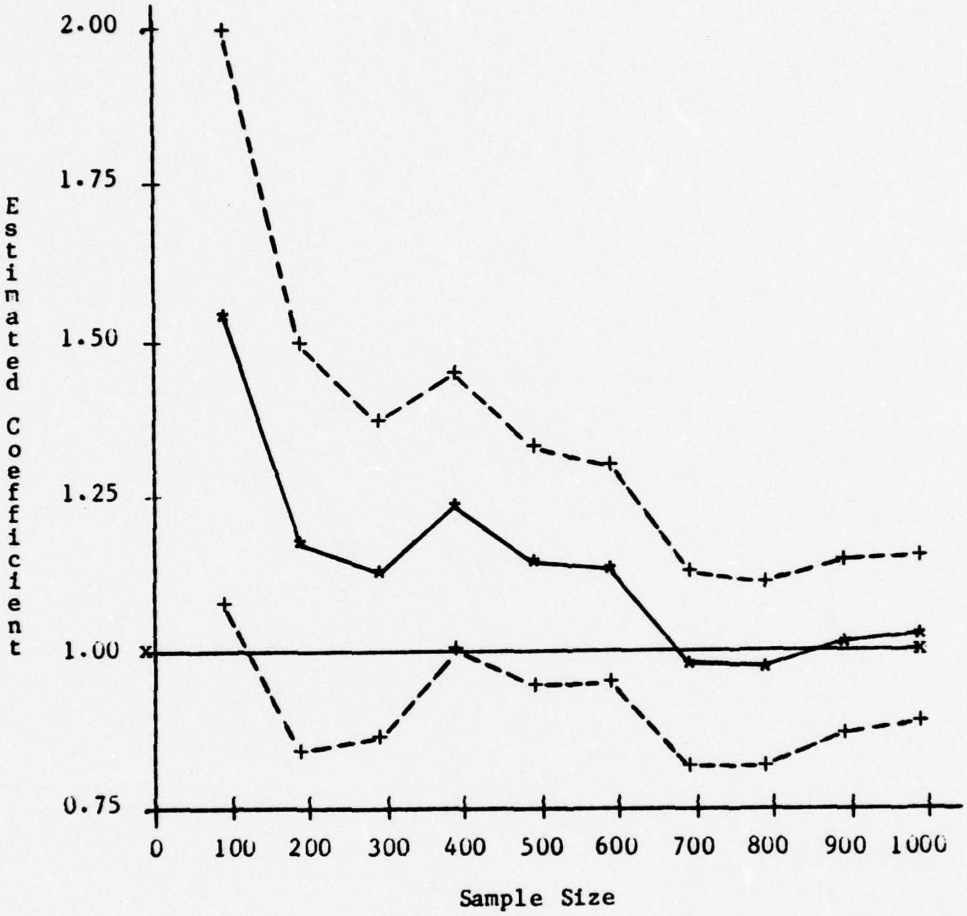


Figure 7: Estimated Beta-4 as a Function of Sample Size (Model A)
(Estimated Standard Error Also Shown)

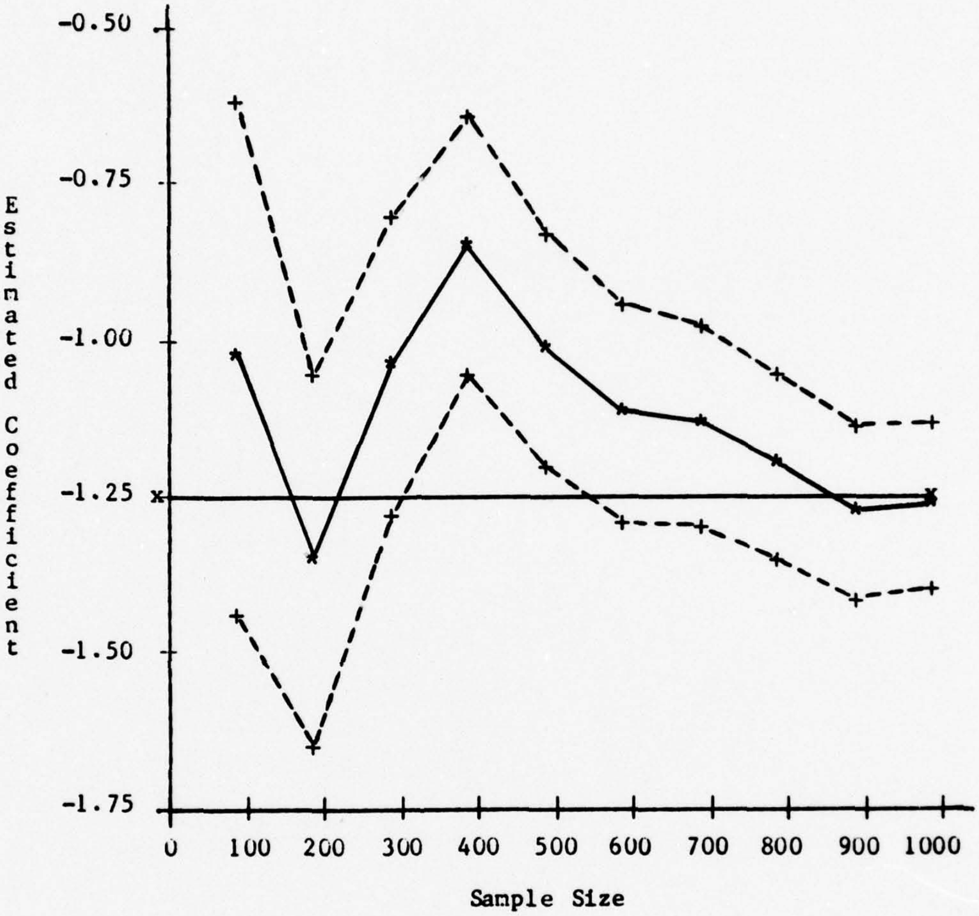


Figure 8: Estimated Beta-5 as a Function of Sample Size (Model A)
(Estimated Standard Error Also Shown)

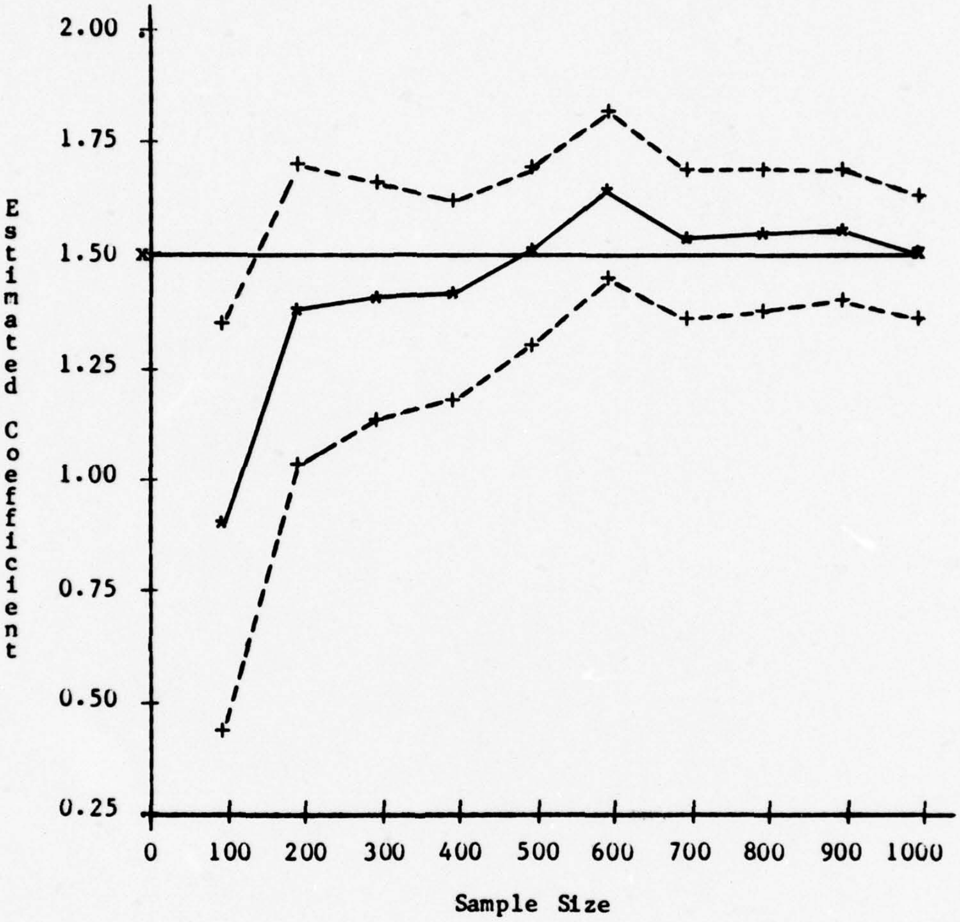


Figure 9: Estimated Beta-6 as a Function of Sample Size (Model A)
(Estimated Standard Error Also Shown)

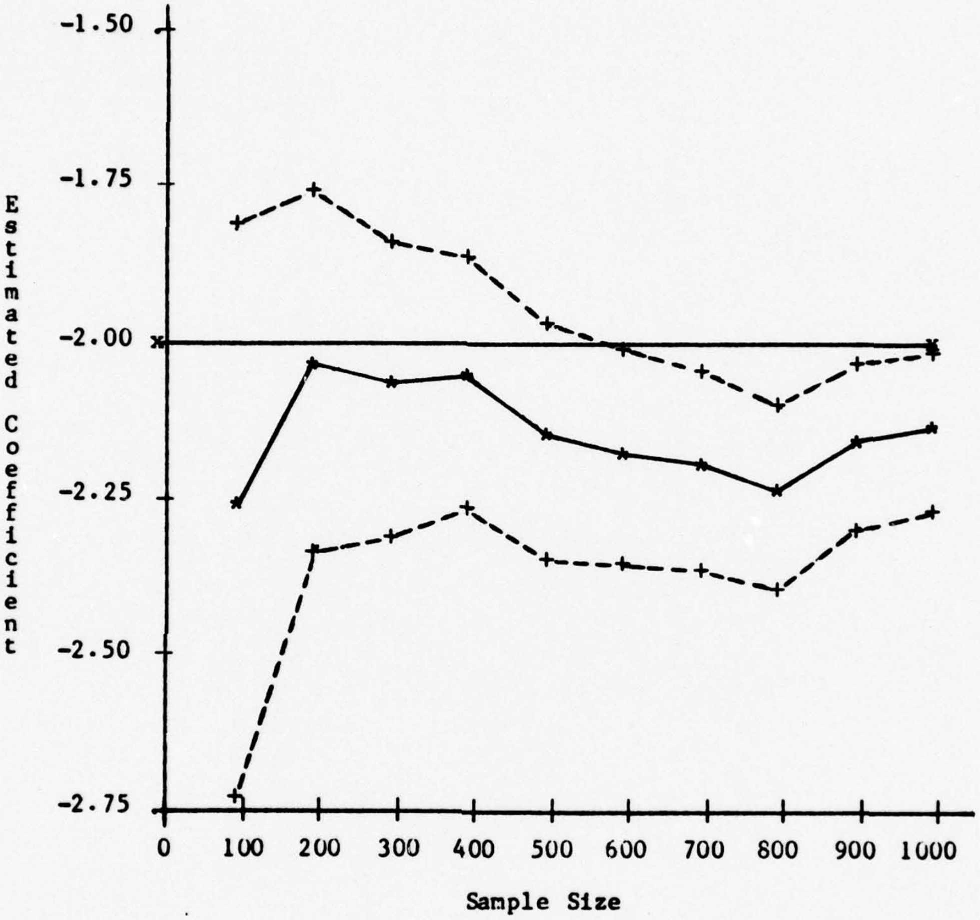


Figure 10: Estimated Beta-0 as a Function of Sample Size (Model B)
(Estimated Standard Error Also Shown)

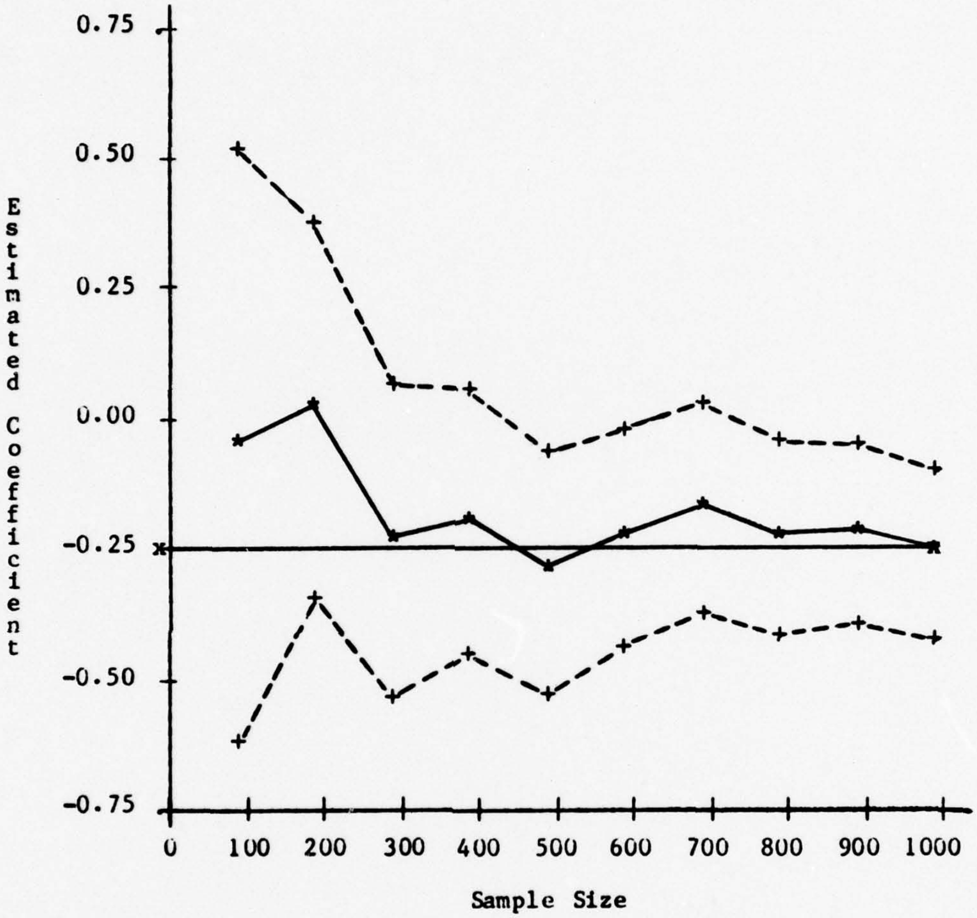


Figure 11: Estimated Beta-1 as a Function of Sample Size (Model B)
(Estimated Standard Error Also Shown)

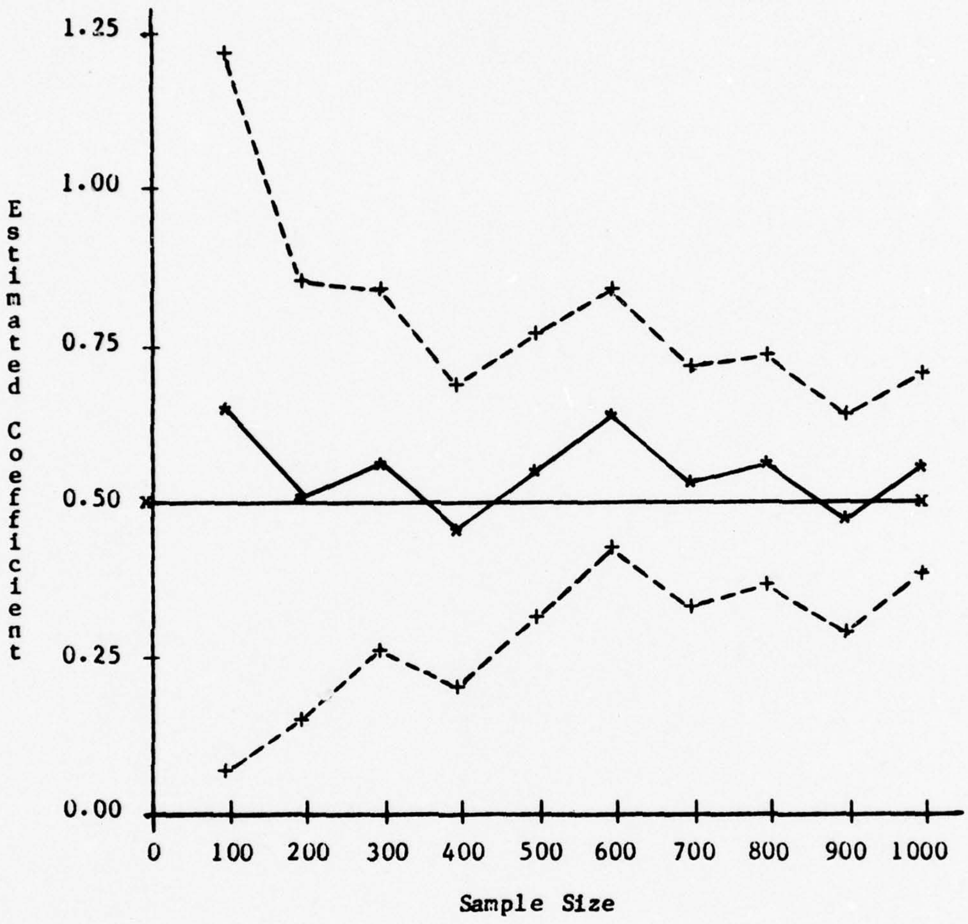


Figure 12: Estimated Beta-2 as a Function of Sample Size (Model B)
(Estimated Standard Error Also Shown)

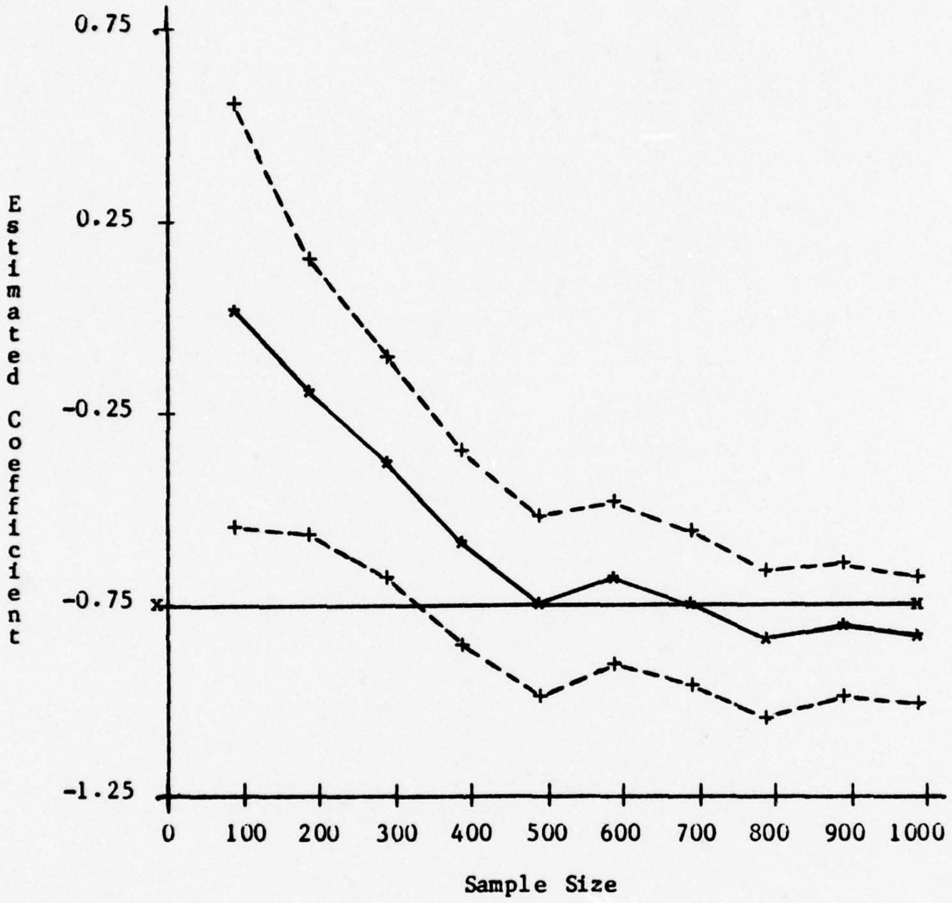


Figure 13: Estimated Beta-3 as a Function of Sample Size (Model B) (Estimated Standard Error Also Shown)

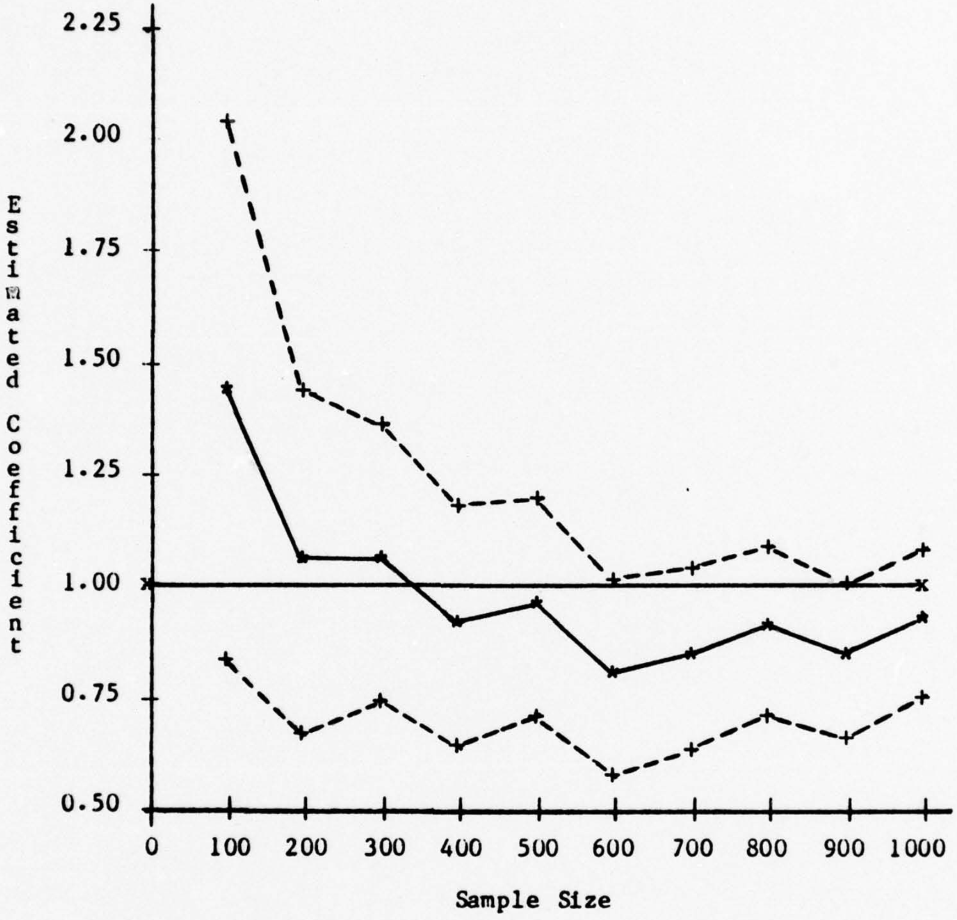


Figure 14: Estimated Beta-4 as a Function of Sample Size (Model B)
(Estimated Standard Error Also Shown)

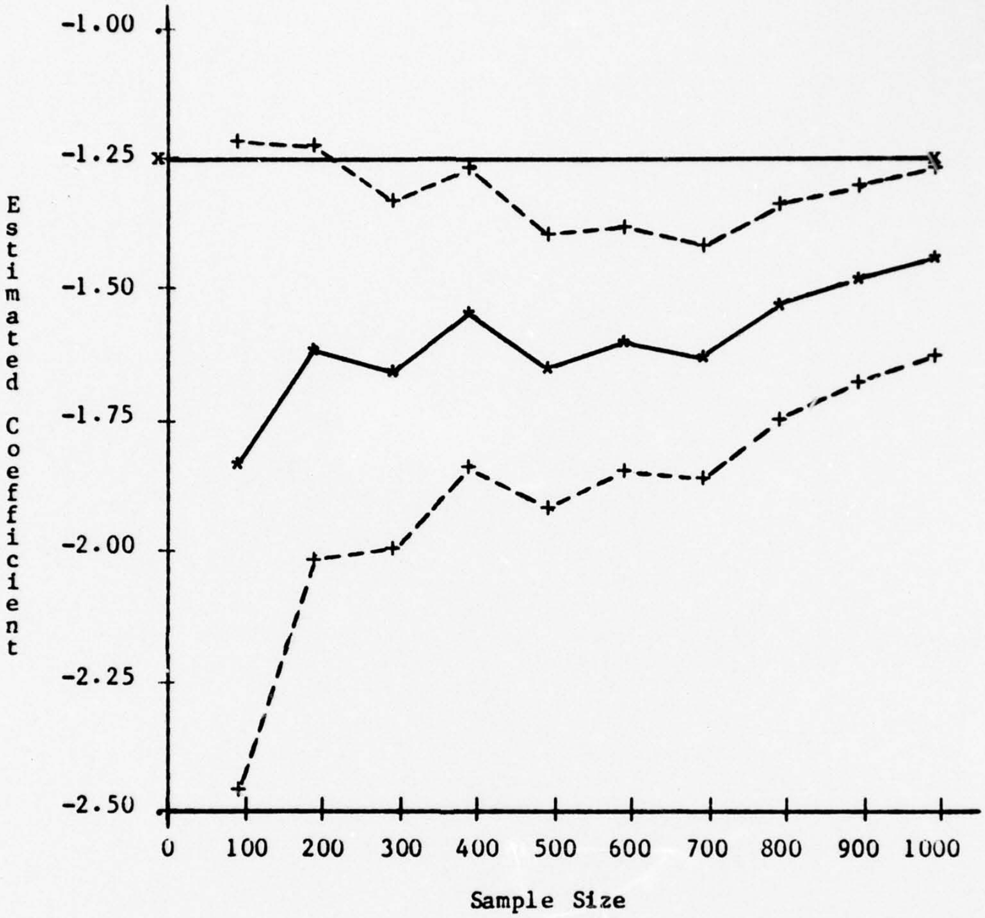


Figure 15: Estimated Beta-5 as a Function of Sample Size (Model B)
(Estimated Standard Error Also Shown)

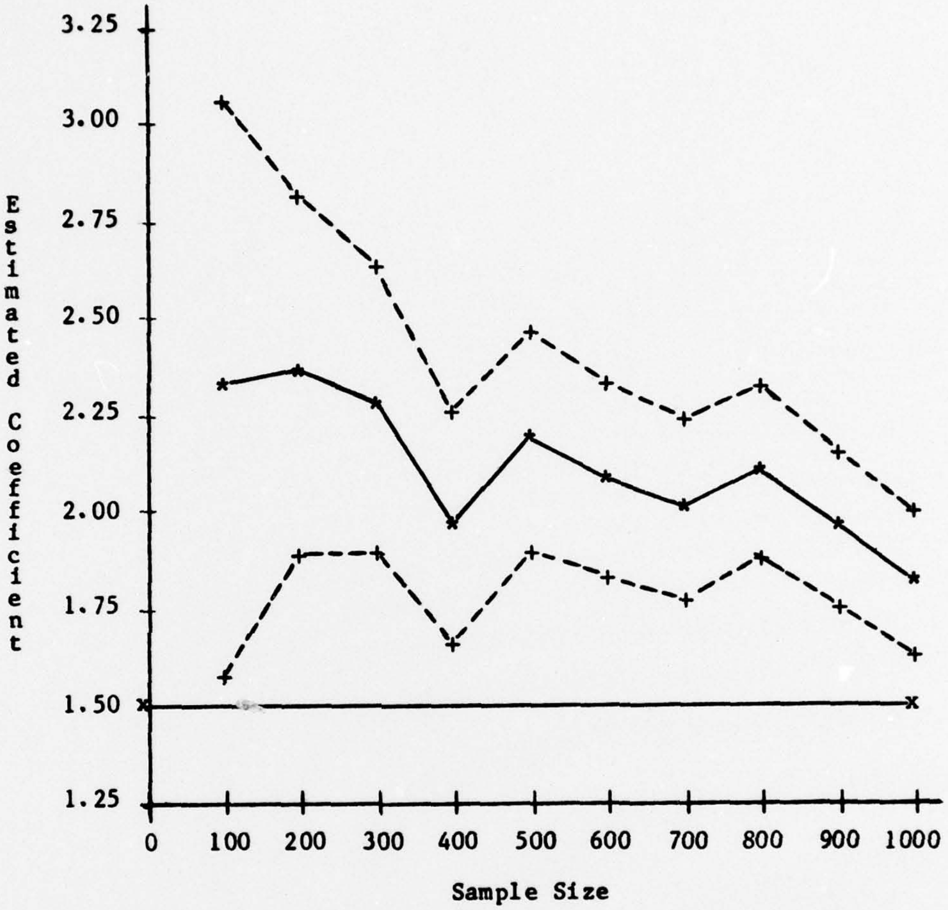


Figure 16: Estimated Beta-6 as a Function of Sample Size (Model B) (Estimated Standard Error Also Shown)

some measure of accuracy, a more useful measure results from a comparison of predicted and true probabilities. To provide a common basis of comparison across sample sizes, each set of estimated coefficients was used in conjunction with equation (1) to derive predicted probabilities for the first 100 observations. Each set of estimated probabilities was paired with the corresponding set of true probabilities, and a linear regression equation was fitted to the data using a weighted least squares procedure.

The results are summarized in Figure 17, which tabulates the intercept, slope and estimated standard error about the regression line, taking the weights into account. In general, the slope of the regression line is near one, the intercept is near zero, and the estimated standard error becomes smaller as sample size increases. The improvement in prediction with increasing sample size can be more dramatically seen by comparing plots of estimated versus true probabilities for various sample sizes.

If the estimated probability prediction model were working correctly, predicted and true probabilities would be expected to cluster about a 45° line between (0,0) and (1,1), and in fact they do. Furthermore, with increased sample size, the clustering about the 45° line becomes tighter. This can be seen in Figures 18 through 23, which correspond to sample sizes of 100, 500, and 1000 for Model A and Model B.

	Sample Size	Intercept	Slope	Standard Error
Model A	100	-0.01	0.92	0.24
	200	-0.03	1.03	0.14
	300	0.00	0.99	0.11
	400	0.02	0.97	0.14
	500	0.00	0.99	0.08
	600	-0.01	1.01	0.08
	700	0.00	0.99	0.06
	800	0.00	1.00	0.05
	900	-0.02	1.01	0.03
	1000	-0.01	1.01	0.03
	∞	0.00	1.00	0.00

	Sample Size	Intercept	Slope	Standard Error
Model B	100	-0.01	1.14	0.20
	200	-0.01	1.17	0.16
	300	-0.01	1.18	0.13
	400	-0.01	1.10	0.08
	500	-0.01	1.16	0.09
	600	-0.01	1.10	0.11
	700	-0.01	1.08	0.09
	800	-0.01	1.08	0.09
	900	-0.01	1.06	0.08
	1000	-0.01	1.05	0.06
	∞	0.00	1.00	0.00

Figure 17: Intercept, Slope, and Standard Error of Regression Line (Predicted Probability versus True Probability)

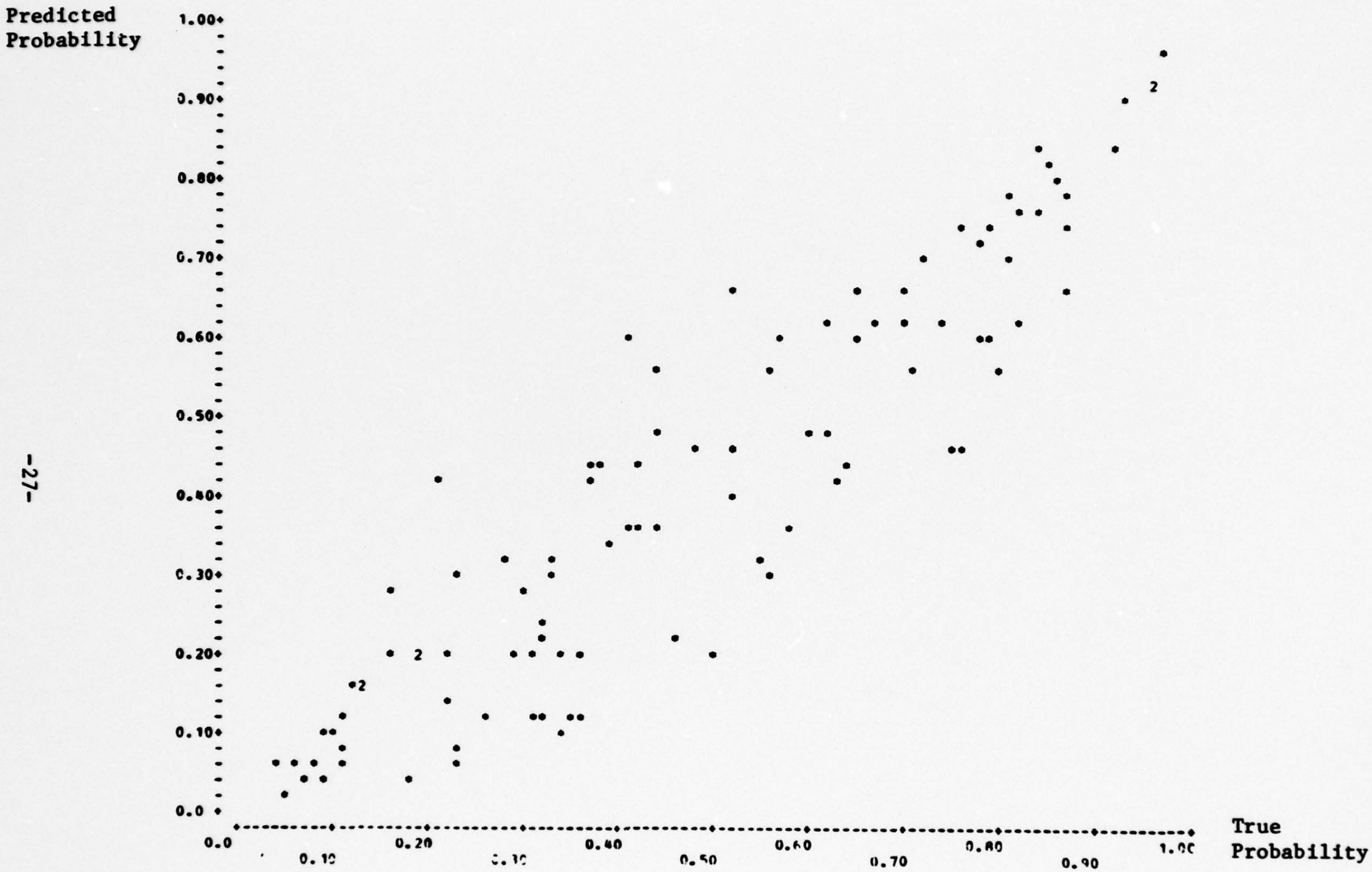


Figure 18: Comparison of Predicted and True Probabilities for a Sample Size of 100 from Model A

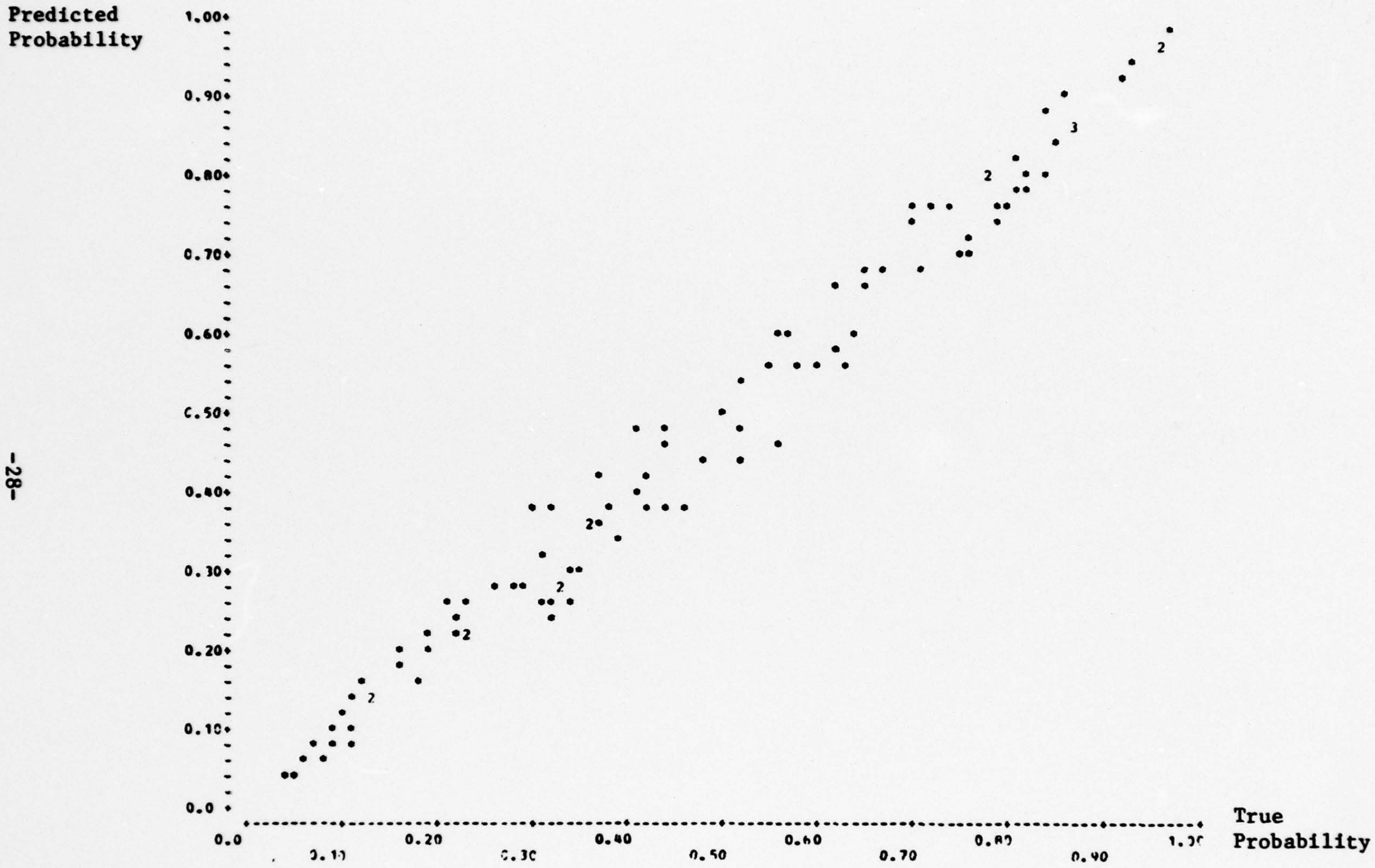


Figure 19: Comparison of Predicted and True Probabilities for a Sample Size of 500 from Model A

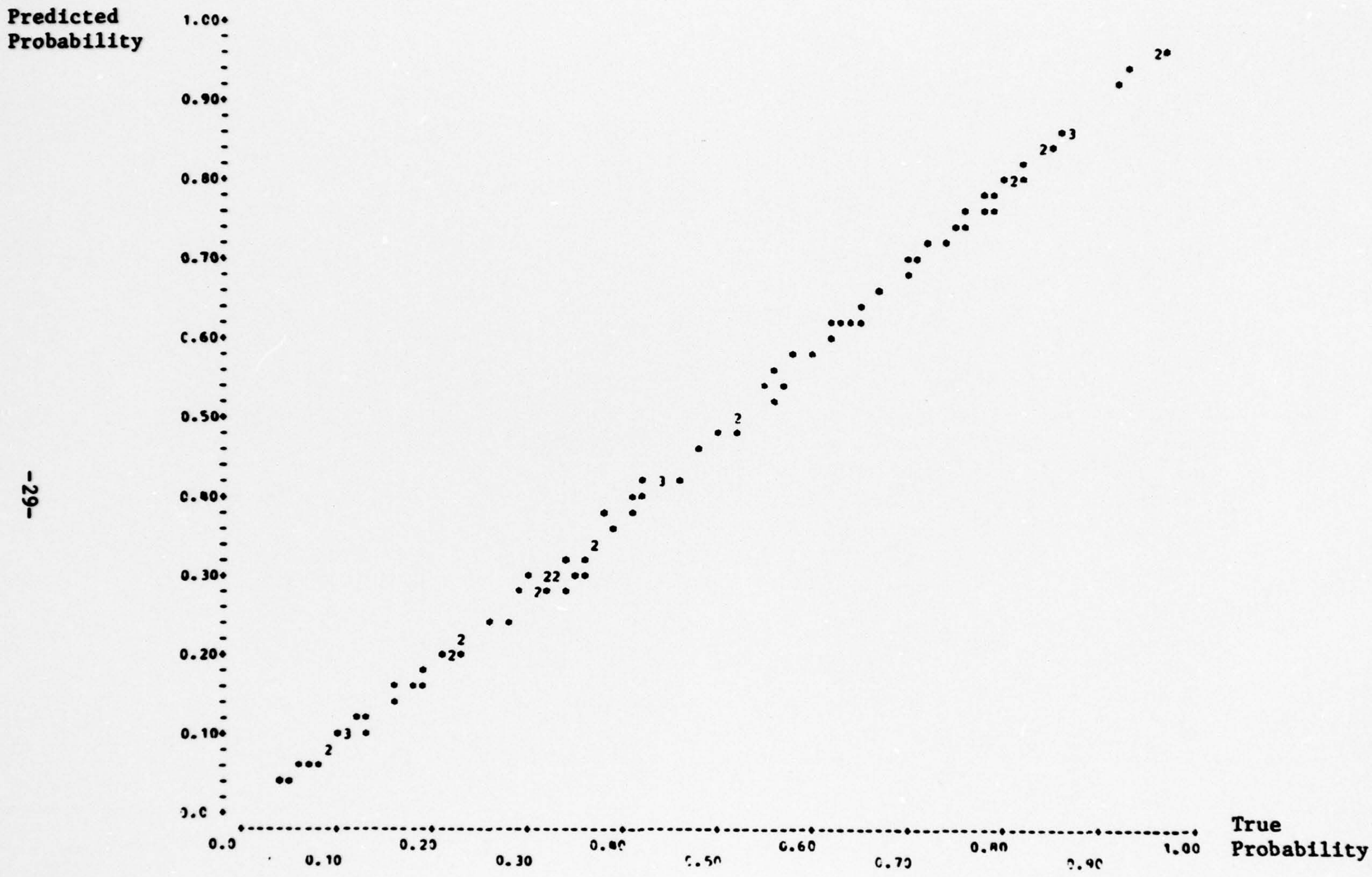


Figure 20: Comparison of Predicted and True Probabilities for a Sample Size of 1000 from Model A

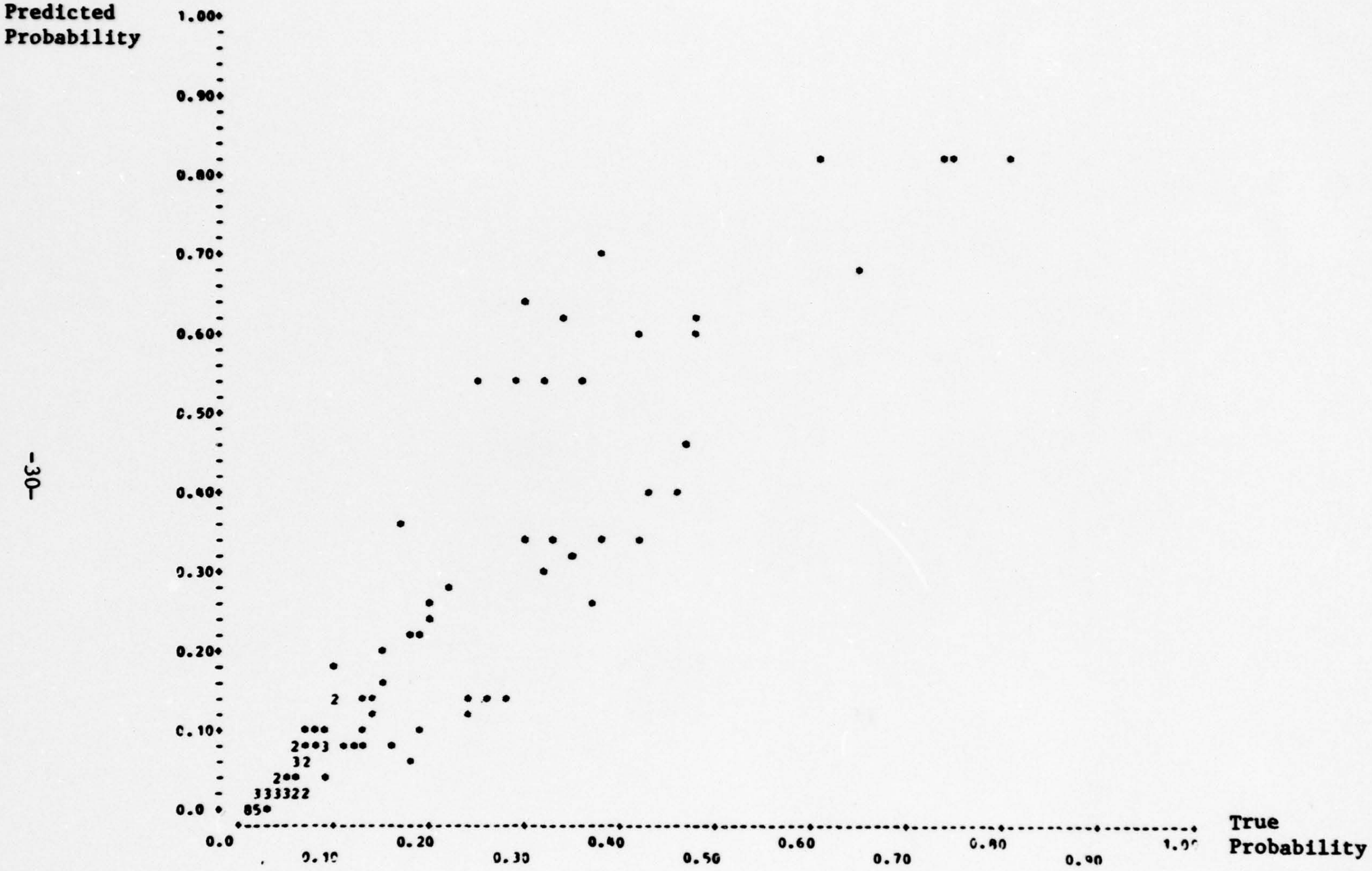


Figure 21: Comparison of Predicted and True Probabilities for a Sample Size of 100 from Model B

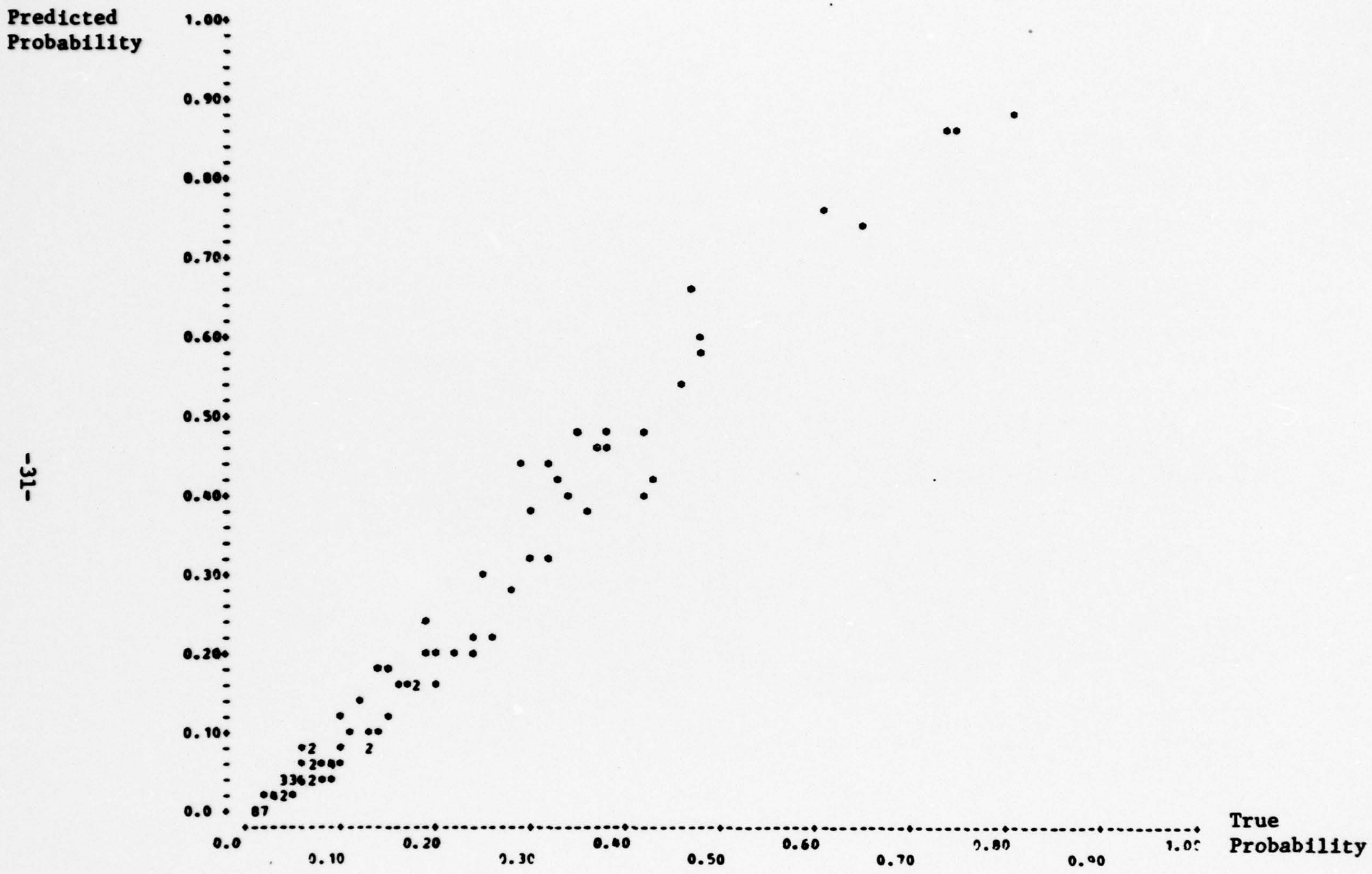


Figure 22: Comparison of Predicted and True Probabilities for a Sample Size of 500 from Model B

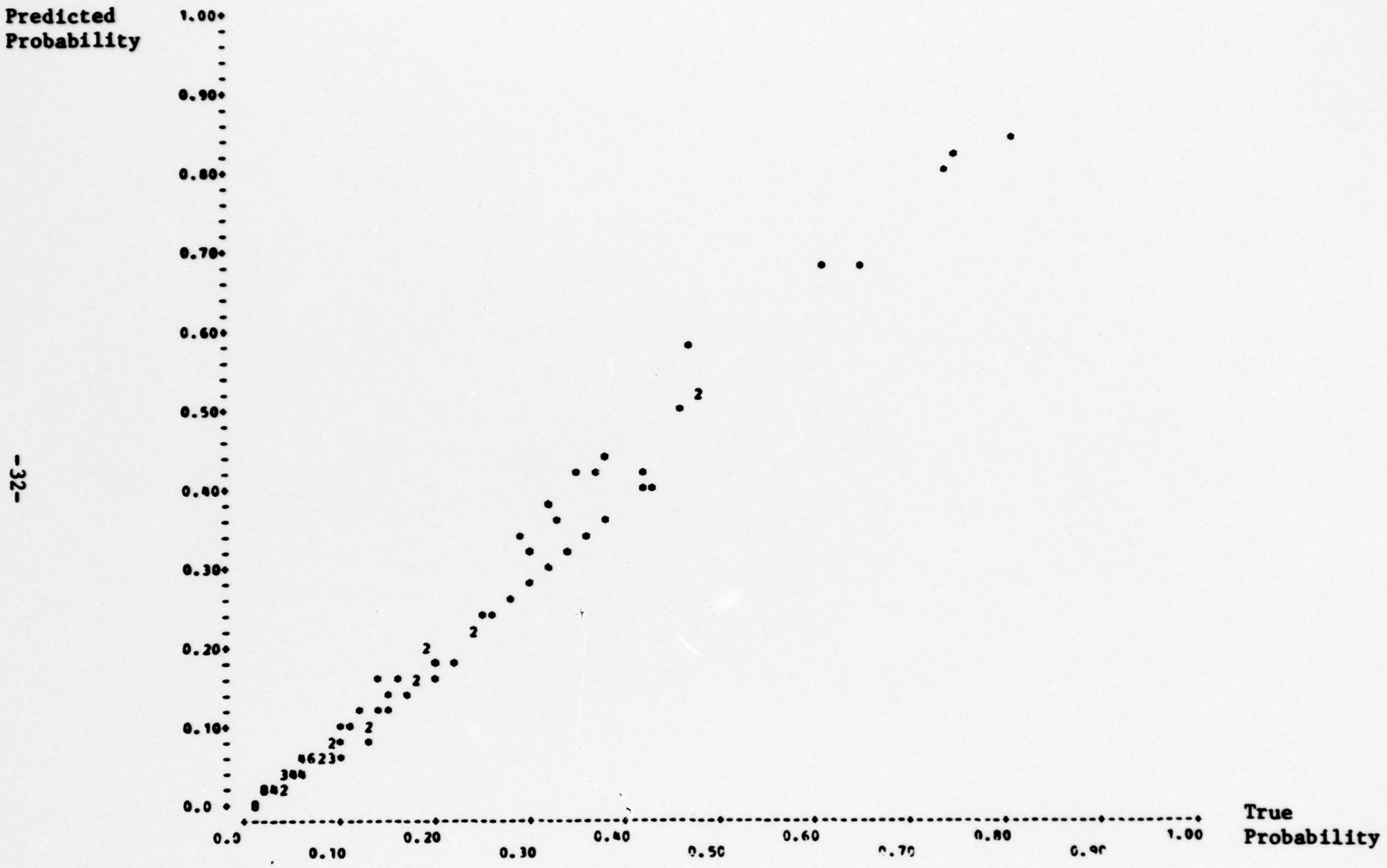


Figure 23: Comparison of Predicted and True Probabilities for a Sample Size of 1000 from Model B

IV. DISCUSSION

Based on the Monte Carlo study discussed in this report, it appears that the procedure used for estimating the coefficients of the logistic model works well. This procedure provides convergence to the correct coefficient values and true probabilities as the sample size increases. Thus, if injury probability as a function of head dynamic response variables may be approximated by a logistic function, the estimation procedure described in this report will yield a satisfactory approximation to that function.

Nonetheless, a major question remains. That question refers to sample size requirements. Of course, larger samples tend to result in better estimates. Also, from the results of this study, it can be seen that overall probability predictions for Model A were better, in general, than those for Model B. This is not surprising, since it would be expected that the best discrimination would result in a data region where the split between occurrences and nonoccurrences was close to 50%-50%.

In general, then, no strong conclusions can be made about required sample size. However, it will be noted that, in both Model A and Model B, agreement between estimated and true probabilities is reasonable even for a sample size of 100, particularly for low probability values. Although there is interest in the overall agreement between estimated and true probabilities, low probability values would constitute major interest. This is because it is desired to exclude those dynamic response conditions for which the probability of injury is greater than some specified small value (for example, 1%, 5%, or 10%). Planned future research will explore, in detail, estimation accuracy in regions of low probabilities.

V. REFERENCES

[1] Jones, R. H., "Probability Estimation Using a Multinomial Logistic Function", J. Statist. Comput. Simul., Vol. 3, pp. 315-329 (1975).

[2] Smith, D. E., "Research on Construction of a Statistical Model for Predicting Impact Acceleration Injury", Technical Report No. 102-2, Desmatics, Inc., 1976.

[3] Walker, S. H. and Duncan, D. B., "Estimation of the Probability of an Event as a Function of Several Independent Variables", Biometrika, Vol. 54, pp. 167-179 (1967).

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER 102-5	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A STUDY OF ESTIMATION ACCURACY WHEN USING A LOGISTIC MODEL FOR PREDICTION OF IMPACT ACCELERATION INJURY		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Dennis E. Smith Robert L. Gardner		8. CONTRACT OR GRANT NUMBER(s) N00014-74-C-0154
9. PERFORMING ORGANIZATION NAME AND ADDRESS Desmatics, Inc. P. O. Box 618 State College, PA 16801		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 207-037
11. CONTROLLING OFFICE NAME AND ADDRESS Biophysics Program (Code 444) Office of Naval Research Arlington, VA 22217		12. REPORT DATE March 1978 13. NUMBER OF PAGES 37
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this report is unlimited.

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Impact Acceleration Injury
Empirical Injury Prediction Model
Statistical Model
Estimation Accuracy Evaluation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report addresses the topic of estimation accuracy in the development of an empirically-based logistic model for predicting impact acceleration injury. Two items of central interest are the degree of accuracy which may be expected for predictions derived from a model and the sensitivity of such predictions to sample size. A Monte Carlo simulation study was undertaken to provide information relating accuracy to sample size for selected model configurations. Two specific sets of model parameters were considered, Monte

→ next page

Carlo samples of various sizes were generated for each, and the accuracy of the resultant predictions were evaluated with respect to the true probabilities.

Based on this Monte Carlo study, it appears that the procedure used for estimating the coefficients of the logistic model works well. This procedure provides convergence to the correct coefficient values and true probabilities as the sample size increases. Thus, if injury probability as a function of head dynamic response variables may be approximated by a logistic function, the estimation procedure described in this report will yield a satisfactory approximation to that function.